Housing Price Prediction for NYC

Yanchao Li , Muyuan Ma, Stanley Li, Qi Lin

Source [1]

NYU

Contents

1. Abstract

With the dataset of published housing property transactions and crime information in New York City, this research develops a house price prediction model at the zip code level by analyzing data via three machine learning models including XGBoost Regression, Random Forest, and Linear Regression machine learning methods. This research also delineates the data processing strategies in detail , and identifies the most influential features for understanding property price.

Key Word

Property Price, Machine Learning models, prediction, NYC

2. Introduction

New York City has one of the hottest real estate markets in the world. One of the most important factors in real estate is its price. Unfortunately, unlike other commodities, the price of real estate does not have a clear mechanism, and their own different factors jointly affect the price change. For example, although a big factor affecting price is its location, different properties in the same area can often vary widely in price. This article hopes to use machine learning to study the intrinsic properties of these houses, and establish a reliable model to analyze their pricing mechanism.

3. Data

3.1 NYC Property Sales (Shape: 99095, 21)

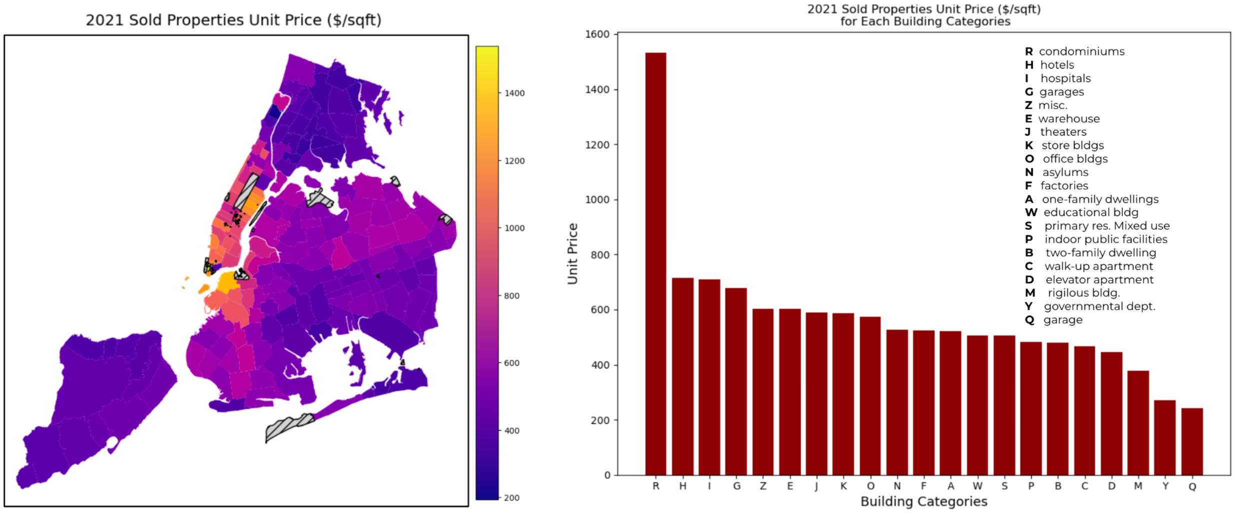
'NYCsales' contains attributes for all properties sold. It is the basis and the most important data source for the follow-up analysis of this paper. It displays information on all transacted properties in New York City. It reveals information such as the year the house was built, zip code, building category, etc, and can be used to find the number of sales and price differences between different areas and building categories.

Data processing include three steps:

1. Delete blank data

2. Delete abnormal sales prices (including prices that are too low or non-positive)
3. Delete unnecessary text information (including text addresses, because this article analyzes in areas of zip codes)

Figure 1 . Left: Number of sales map;
Right: Difference in the average sale price of building categories



Note: Building categories see Appendix B.

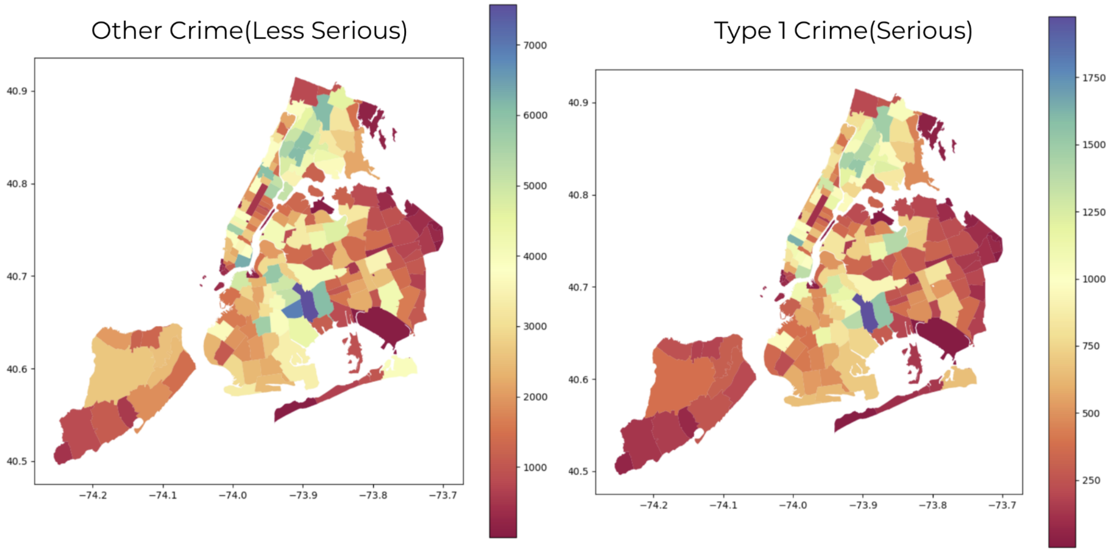
3.2 NYC Crime Cases(Shape: 7825499,35)

'Police' contains case statistics for 18 common crime types in New York City. It was added as an attribute to the 'NYCsales' dataset participation study in subsequent analyses. It exists to describe the state of law and order in the neighborhood where the property is located.

Data processing include following steps:

1. Delete blank data
2. Delete abnormal case time (only keep 1/1/2021-1/1/2022)
3. Obtain the geographical coordinates of the case and assign the zipcode where the coordinates sit
4. Divide the data into 'Type1' (serious crime) and 'Other crime' according to crime type
5. Count the above two types of crimes according to the zip code and add them as features to 'NYCsales'

Figure 2 . Left:Other Crime map;   Right: Type 1 Crime map



4. Methodology

4.1 Random Forest Regression

Random forest is a combination of tree predictors. Each tree depends on the value of a random vector sampled independently and with the same distribution for all trees in the forest. When the number of trees in the forest becomes large, the generalization error a.s. of the forest converges to a limit. Random forest is an improvement of the bagging method. It only randomly selects a subset of features and uses the best segmentation feature to split each node in the tree, which is more robust in terms of noise.

4.1.1 Preliminary Model

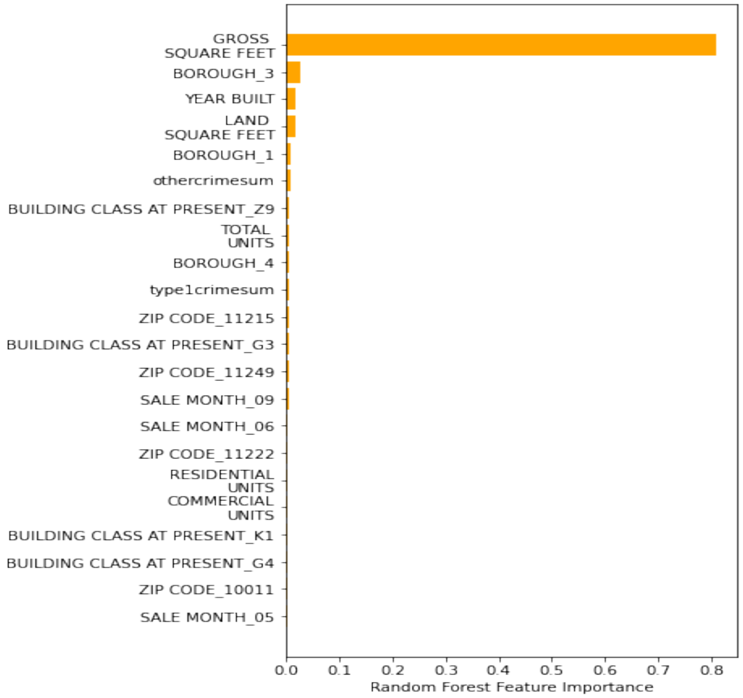
In the preliminary model of the random forest algorithm, we used all variables as input to the model, where the numerical variables were left as they were, and the categorical variables were transformed into dummy variables. 80% of the data set was then used as the training set and 20% as the test set. When training the model, we used the GridSearchCV function to adjust the model's parameters and used 3-fold cross-validation. Table 1 shows the parameters we tuned in the preliminary model, along with their tuning ranges and optimal values. The total run time of the baseline model is 56min 58s.

Table 1. Grid search hyper-parameters tuning for the preliminary model of random forest

Hyper-parameters	Searched	Selected
max_depth	[20, 40, 60, 80, None]	40
min_samples_leaf	[1, 2, 4]	1
min_samples_split	[2, 5, 10]	2
n_estimators	[200, 400, 600, 800]	200

After tuning the model, we used the test set's performance to validate the model's accuracy. This preliminary model achieved RMSE and R2 values of 468.16 and 0.33 on the test set. Next, we use the feature importance in this model to select the variables that impact the housing price per unit area. After sorting the variables according to the value of feature importance from high to low and adding them together, the sum of the feature importance of the first three variables exceeds 0.85. And the sum of the first nine variables exceeds 0.90, and the sum of the first twenty-two variables more than 0.95. Figure 3 shows these twenty-two variables and their feature importance.

Figure 3 . Feature Importance of Random Forest Model



4.1.2 Improved Model

After transforming categorical variables into dummy variables, the model has more than three hundred variables. Still, the inclusion of most of them does not significantly improve the interpretability of the model. Therefore, in the improved model, we selected only the twenty-two variables with the highest feature importance as inputs to the model and modeled them again with the random forest algorithm. In this model, we used the same training set test set splitting approach and the exact tuning of the parameters as in the preliminary model. Table 2 shows the parameters adjusted for the model with the selected variables and their tuning ranges and optimal values. The total run time of the improved model is 14min 20s.

Table 2. Grid search hyper-parameters tuning for the improved model of random forest

Hyper-parameters	Searched	Selected
max_depth	[20, 40, 60, 80, None]	80
min_samples_leaf	[1, 2, 4]	2
min_samples_split	[2, 5, 10]	2
n_estimators	[200, 400, 600, 800]	200

We then validated the model's accuracy on the test set. The model achieved RMSE and R2 values of 457.83 and 0.36, respectively. By comparing with the baseline model, it can be found that although the performance of the model still needs improvement, the improved model achieves better performance on the test set and also saves a lot of running time.

4.2 Extreme Gradient Boost (XGBoost) Regression Model

XGBoost Regression is one of the ensemble machine learning models. The ensembles are constructed from decision tree models, with trees added one at a time to the ensemble and fit to correct the prediction errors made by prior models. XGBoost was implemented to maximize the efficiency of computing time and memory resources.

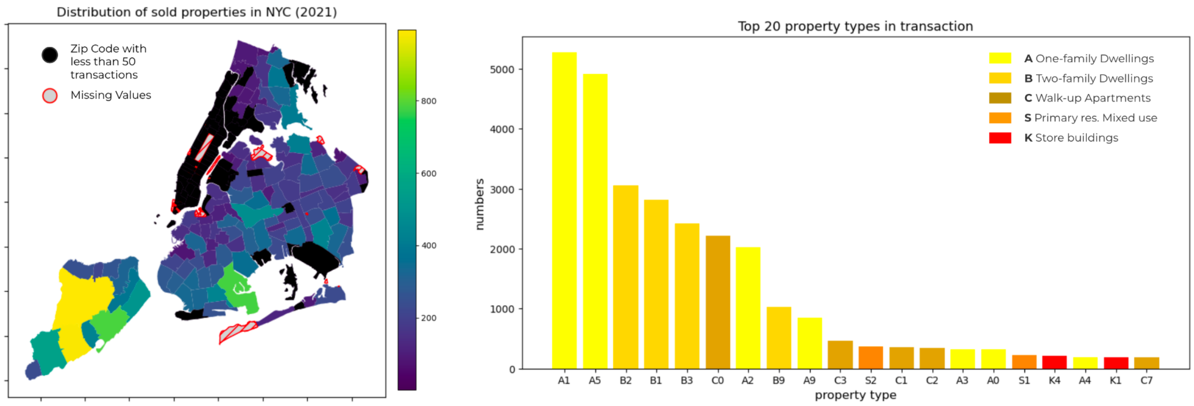
4.2.1 Preliminary Model

29,538 observations with 312 features were used to fit the XGBoost regression model with its default parameters. We use the result of ten-fold cross validation mean rooted square error of 190.47 to measure the preliminary model performance. In terms of the feature importance, we found the top ten most important features are spatial features (i.e., borough location of property) and environment safety features (i.e., yearly crime cases). 57 of the 312 features are detected to have no feature importance to understanding the whole data.

4.2.2 Improved Model

We took three steps to improve the model. We removed the 57 non-important features mentioned above and checked the data quality to filter out potential misleading features. After that, we tuned the XGBoost model to find the best parameter for minimizing (rooted) mean squared prediction errors. We mainly checked two categorical features in the data quality check step, “zip code” and “building category.” As mentioned in the data processing part, housing unit prices vary in dimensions. In the zip codes of Manhattan, the mean property transactions in 2021 are generally less than 50, while the most frequent transactions happened in the zip codes of Staten Island, as shown in figure 4 (a). In sold property categories, figure 4 (b) indicates that the top 20 popular properties are residential buildings (i.e., one-family, two-family, walk-up apt., and mixed-use with primary residential use.) and small-size stores. These transactions account for 95.8% of total transactions. Building on this analysis, we select features of zip codes with transactions of more than 50 (representing 97.5 total transactions) and the top 20 building categories.

Figure 4 . Feature selection (a) Distribution of Sold Properties in NYC (2021), (b) Top 20 Property Types in Transaction

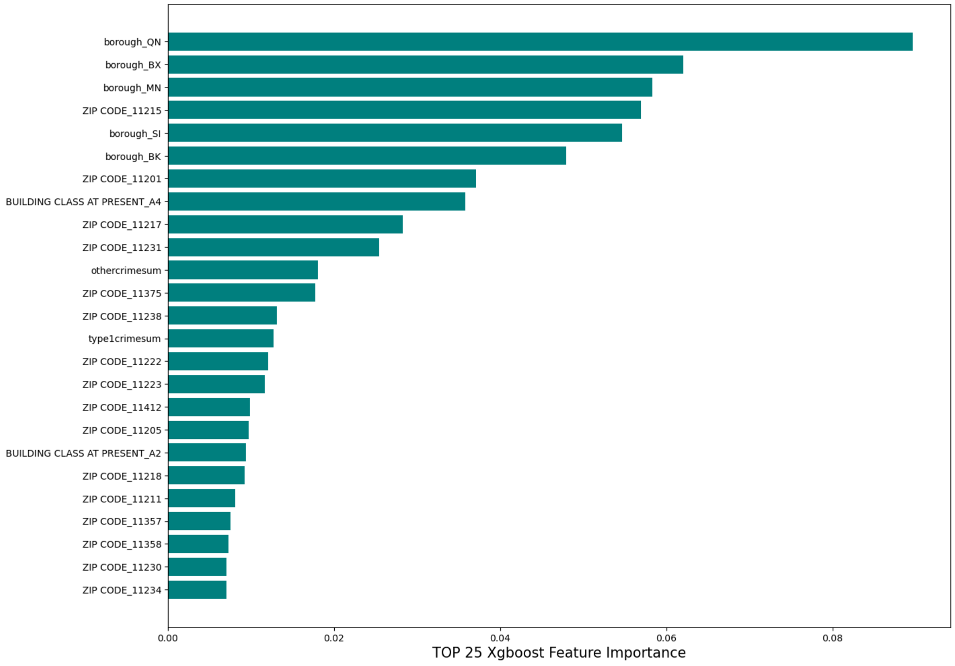


After selecting the remaining 127 features, we run the XGBoost model with a parameter space to find the best parameter set. Table 3 shows the best parameter set (bolded) among the spaces. With the three improvement steps, the prediction of rooted mean squared error reduced to 175.03, indicating better performance. The most influential features related to unit price are property location, current building classes and environmental safety, see Figure 5 for details.

Table 3. Hyperparameters Tuning for the Improved XGBoost Regression Model

Hyperpameters	Value Ranges and the Selected (bolded)
colsample_bylevel	[.1, .2, .3, .4, .5, .6, .7, .8, **0.9**, 1]
colsample_bytree	[.1, .2**, 0.3**, .4, .5, .6, .7, .8, .9, 1]
learning_rate	[0.01,0.02,0.05,**0.1**,0.2,0.5,1]
max_depth	[3,**4**,5,6,7,8,9,10]
min_child_weight	[**1**,3,5,7]
n_estimators	[100,200,300,400,**500**,1000]
reg_alpha	[20,30,32**,34**,35,36,38,40,50]
reg_lambda	[0.1,0.3,0.5,0.7,0.9,**1**]
subsample	[0.5, 0.6, 0.7, 0.8, **0.9**, 1]

Figure 5 . Feature Importance of XGBoost Regression Model



4.3 Linear Regression

Linear regression is a statistical method used to model the linear relationship between dependent and independent variables. The strength of linear regression is that it is a relatively simple method that can be easily implemented and understood. However, it assumes a linear relationship between the dependent and independent variables, which may only sometimes be the case.

4.3.1 Preliminary Model

We began by normalizing the variables in our dataset. After normalizing the variables, we fit a complete model as a baseline. This gave us a starting point for comparison and allowed us to see how the model performed before any feature selection was applied.

4.3.2 Improved Model

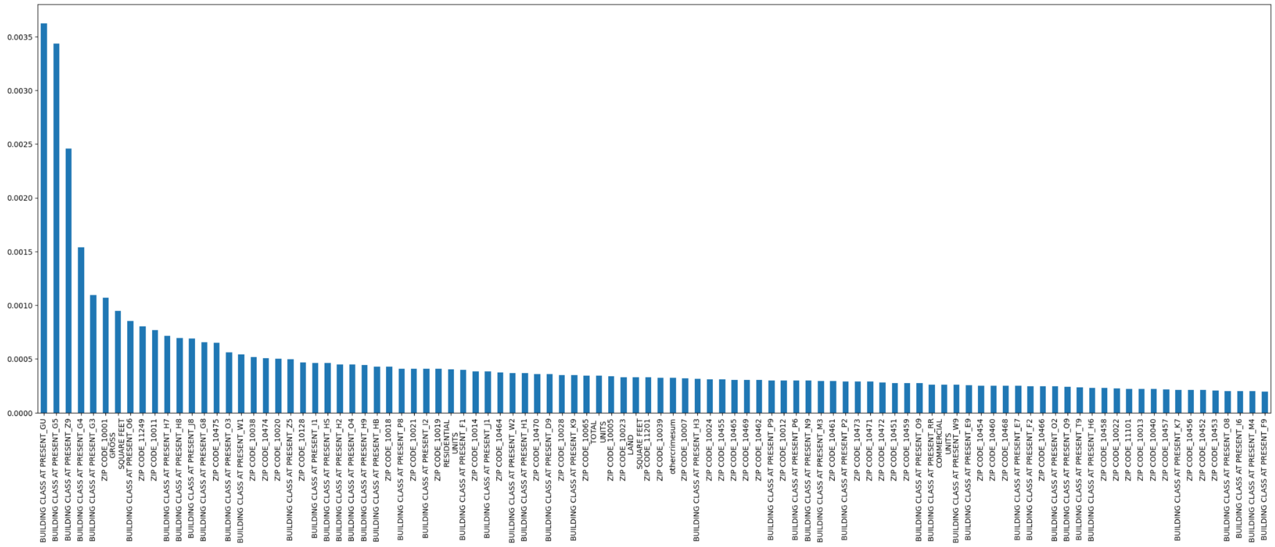
We used variance inflation factors (VIFs) to identify and remove highly correlated features. By removing correlated features, we aim to improve the stability and interpretability of the model. After removing correlated features, we applied four methods to select a subset of the most relevant features. The first method was to obtain importance from coefficients. Features with the highest weights were considered the most important, and we selected a subset of the top 10 features based on this criterion. The second method involved removing variables with a high p-value. We removed all variables with a p-value above 0.01 and selected a subset of 182 variables based on this criterion. The third method is forward selection using Akaike Information Criterion (AIC) criteria. We used AIC to select the best model among a set of candidates and selected a subset of 15 variables based on this criterion. The fourth is the recursive method, which removes features one at a time and chooses the model with the highest performance. We used a cross-validation procedure to ensure that the chosen model generalizes well to unseen data. By selecting a subset of the most important features, we aim to reduce overfitting, improve the stability of the model, and make the model more interpretable. Then, we fit a new regression on selected features. However, all methods used needed a better R-squared score. Table 4 shows the performance of each regression model after feature selection.

Table 4. Feature Selection Methods Comparison

	Coefficient (Normalized)	P value <0.01	Forward Selection	Recursive Feature Elimination
R2 Score	0.23	0.25	0.25	0.18
Number of varibles	10	182	15	15

Table 4 also indicates that the second method (using p-values to select features) had the highest R-squared score. The recursive feature elimination method had the lowest R-squared score, indicating that it may not be as effective as the other methods. Still, the resulting model may need to be more interpretable. The coefficient method and the forward selection method had the same R-squared score, but the coefficient method included fewer variables, making it potentially more interpretable. Figure 6 illustrates the top 50 components with the high absolute value of coefficients, indicating their importance in the model. Therefore, we decided to use the features selected by the coefficient method for the regression analysis.

Figure 6 . Top 50 Variables with High Absolute Value of Coefficient



5. Result

5.1 Model Performance Comparison

Compared with ensemble machine learning methods of Random forest and XGBoost, the linear regression model better interprets the variables. With the insights into data structure derived from our two ensemble methods, we improved the linear regression model to achieve a balance of prediction and interpretation, as shown in Table 5.

Table 5. Model Performance Comparison

Models	Rooted Mean Square Error	Out of Sample R-squared
Random Forest Regression	457.83	0.36
XGBoost Regression	175.03	0.48
Linear Regression	196.93	0.34

5.2 Property Prediction Formula

Building on our previous ensemble model results, we improved the explanation of our linear model to complete the property prediction formula. The improvement strategies for linear regression model include selecting special features of “zip code” and “building categories” (as shown in the section of the XGBoost model) and binning features of “month” and “built year.” After these steps, the model performed better in capturing the data, with a root mean square error sharply decreased to 196.63. As shown in Table 6, we can find detailed property prices rules. For example, compared with property sold in Brooklyn, the average unit price normally would be $ 228.3 higher in Manhattan if other conditions remain unchanged.

Table 6. Improved Linear Regression Model Result

Column Header 1	COEF.	P-VALUE
CONST	462.36	0.000
RESIDENTIAL_UNITS	-1.95	0.000
COMMERCIAL_UNITS	1.8	0.398
TOTAL UNITS	-0.15	0.889
TYPE ONE CRIME SUM (per 100)	-54.6	0.000
OTHER CRIME SUM (per 100)	-22.2	0.000
TAX CLASS AT PRESENT_2	-66.8	0.484
TAX CLASS AT PRESENT_2A	-17.56	0.850
TAX CLASS AT PRESENT_2B	-32.5	0.729
TAX CLASS AT PRESENT_4	-27.9	0.000
TAX CLASS AT TIME OF SALE_2	-13.8	0.021

TAX CLASS AT TIME OF SALE_4	20.3	0.890
BORO_BX	-189.8	0.000
BORO_MN	228.3	0.000
BORO_QN	-96.9	0.000
BORO_SI	-203.4	0.000
SEASON_S2	13.7	0.000
SEASON_S3	31.8	0.000
SEASON_S4	33.2	0.000
BLDG_CAT_B	-67.5	0.000
BLDG_CAT_C	-122.5	0.000
BLDG_CAT_K	27.9	0.705
BLDG_CAT_S	-138.7	0.000
IS_TOPZIP_Y	455.1	0.000
IS_BOTZIP_Y	-33.5	0.000
IS_TOPYEAR_Y	82.4	0.000
IS_BOTYEAR_Y	-70.3	0.000

Note: Base model: BORO_BR, SEASON_S1, CLDG_CAT_A, IS_TOPZIP_N, IS_BOTZIP_N, IS_TOPYEAR_N, IS_BOTYEAR_N; Column explanation see Appendix B.

5.3 Application

The linear regression model result of our research can be used to understand the property unit price in NYC. Specifically, it can be developed into an interactive unit price calculator to predict the zip code-level property price. By asking the user questions shown in Appendix D, users will get a price result, which can then be used as a reference to the real-specific property price.

6. Conclusion

This research focuses on understanding and predicting the real estate market in NYC by integrating cutting-edge sales and environmental safety data. Building on the analysis of three machine learning models, including random forest, XGBoost, and linear regression, the research developed a property price prediction model to enable zip code price prediction.

The research also has limitations. The study employs “zip code” as a categorical variable in predicting house prices. In future research, the authors will incorporate more numerical variables such as median income, education rate, and transportation accessibility to describe the specific characteristics of each spatial zone.

7. Appendix

A. Team Member Contribution

Team member	Contribution
Muyuan Ma	• Dataset Analysis, Processing; • Attempt on Decision Tree, not included due to bad performance • Presentation • Writing
Stanley Li	• Linear Regression Model- First Round • Presentation • Writing
Yanchao Li	• XGBoost Regression Model • Linear Regression Model- Second Round • Presentation • Writing
Qi Lin	• Random Forest Model • Presentation • Writing

B. Building Categories

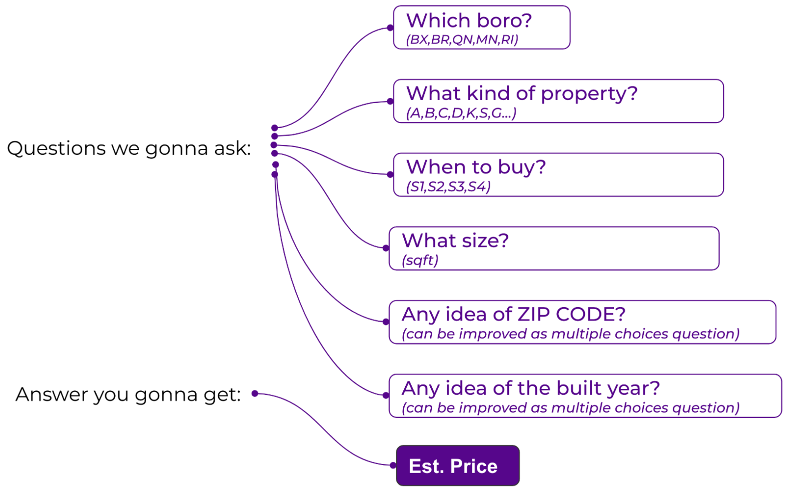
Categories start with	Building categories refer to

A,B,C,D, R,S	Residential types
K,O	Commercial types
(all remaining types, i.e.:) E,F,G,H,I,J,M,N,P,Q,W,Y,Z	Other facilities

C. Original Linear Regression Coefficients

Variables Selected	Coefficients
BUILDING CLASS AT PRESENT_GU	0.00370
BUILDING CLASS AT PRESENT_G50	0.00370
BUILDING CLASS AT PRESENT_Z9	0.0025
BUILDING CLASS AT PRESENT_G4	0.0017
BUILDING CLASS AT PRESENT_G3	0.0012
ZIP CODE_10001	0.001
LAND SQUARE FEET	-0.0009
BUILDING CLASS AT PRESENT_H7	-0.0002
BUILDING CLASS AT PRESENT_O6	-0.0001
ZIP CODE_11249	0.0009

D. Interactive Prediction Questions



Questions we gonna ask:

Which boro?
(BX,BR,QN,MN,RI)

What kind of property?
(A,B,C,D,K,S,G...)

When to buy?
(S1,S2,S3,S4)

What size?
(sqft)

Any idea of ZIP CODE?
(can be improved as multiple choices question)

Any idea of the built year?
(can be improved as multiple choices question)

Answer you gonna get:

Est. Price

9. Reference

IMAGE SOURCE

DATA SOURCE

[1]NYC Sales Data https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page

[2]NYC Crime Cases Data,
https://data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr

PACKAGES

[3]XGBoost Regressor:
https://xgboost.readthedocs.io/en/stable/python/index.html

[4]RandomForestRegressor:
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[5]Linear Regression
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html